Robust Correlation Procedure via *Sn* Estimator

Nor Aishah Ahad, Nur Amira Zakaria, Suhaida Abdullah, Sharipah Soaad Syed Yahaya, Norhayati Yusof
School of Quantitative Sciences, Universiti Utara Malaysia
aishah@uum.edu.my

***Abstract*—Pearson correlation coefficient is the most widely used statistical technique when measuring a relationship between the bivariate normal distribution when the assumptions are fulfilled. However, this classical correlation coefficient performs poor in the presence of an outlier. Therefore, this study aims to propose a new version of robust correlation coefficient based on robust scale estimator S_n. The performance of the proposed robust correlation coefficient is assessed based on correlation value, average bias and standard error. The performance of the proposed coefficient is compared with the classical correlation together with the existing robust correlation coefficient. Classical correlation coefficient performs well under the condition of perfect data. However, its performance becomes worst when data is contaminated. Under the condition of data contamination, robust correlation coefficient performed better compared to classical correlation.**

***Index Terms*—Average Bias; Outlier; Robust Correlation Coefficient; *Sn* Estimatot.**

I. INTRODUCTION

Bivariate normal distribution consists of two random variables that are normally distributed and let indicate it as X_i and Y_i where $i = 1, 2, 3 \ldots n$ is an observation from it. The parameters for the bivariate normal distribution are (μ_x, μ_y, σ_x^2, σ_y^2). The mean of x and y are represented by μ_x and μ_y meanwhile the variance of x and y represented by σ_x^2 and σ_y^2. The correlation coefficient denoted by ρ summarises the association between bivariate data.

Correlation is measured by a statistic called the correlation coefficient, which aims at characterising the strength of the association between two variables. It is a dimensionless quantity that takes a value in the range of -1 to 0 to +1, where no units involved. The strength of the relationship can be anywhere between -1 to +1. The stronger the correlation, the closer the correlation coefficients comes to ±1.

The most frequently used correlation coefficient among practitioners is the Pearson correlation coefficient. This coefficient is very powerful when there is a linear relationship between the two variables and the distribution is normally distributed. However, when there is the existence of outlier, normal distribution usually deviates, and this will reduce the capability of the Pearson correlation coefficient to measure the strength of the relationship. An outlier is an observation in a sample that deviates markedly from the other observation [1]. The distortion that caused by the existence of the outlier tends to mislead the interpretation of the relationship between variables. Thus, the nonparametric method is one of the solutions for this problem. Nonparametric correlation coefficients such as Spearman rank correlation coefficient and Kendall's tau correlation coefficient is the coefficients that are suitable to use under non-normal data. Despite that, these coefficients performance is not as good as the Pearson correlation coefficient when data is normally distributed with the linear relationship because of the usage of rank values instead of the original observations.

To handle the presence of an outlier in the bivariate data, besides using the nonparametric procedure, the robust approach also can be considered. The robust statistical procedures have been promoted as alternatives to solve parametric methods that did not meet the assumptions [2, 3]. Robust correlation coefficients were also developed as options to the Pearson correlation coefficient [4, 5].

II. LITERATURE REVIEW

Pearson correlation is the most widely used as a parametric technique to measure the strength of the relationship between two variables due to its simplicity in the calculation and the excellent performance when data is normally distributed with a linear relationship. However, this coefficient suffers from the existence of outlier [4, 5, 6, 7]. For example, if there is a positive relationship between two variables, this coefficient unable to detect the relationship if there is only one outlier presence. Since Pearson's correlation is sensitive to the outlier, therefore many researchers realise the necessity in robust counterparts of the sample correlation coefficient.

In 1990, a robust correlation coefficient with a high breakdown point based on the least median of squares (*LMS*) regression procedure was proposed [4]. This robust correlation coefficient based on least median squares (*LMS*) as an estimator provide a higher breakdown point than the existed correlation coefficient. However, the *LMS* produced a bad result when there are errors in normally distributed data. It also tends to give unrealistically value for correlation coefficients whether high or low [5]. For instance, if the correlation between two variables exists with the only moderate relationship, the *LMS* correlation coefficient tends to provide the very high value of the relationship. As an alternative to overcome this problem, [4] proposed robust correlation coefficient using weighted least squares by combining the *LMS* estimator with *M*-estimator

During 2011, a new version of robust correlation coefficient based on the median using scale estimator median absolute deviation (*MAD*) was proposed and known as Median-Product (MP) correlation coefficient [8]. They replaced the mean in classical correlation coefficient into the median and used *MAD* in this coefficient calculation. However, *MAD* consists of a few drawbacks. Firstly, this estimator has low efficiency, which is 37% at the Gaussian distribution and secondly, *MAD* only view a dispersion of symmetric distribution [9]. The advantage of this robust correlation coefficient is that it requires less computing time when compared with the existing robust estimators that have been proposed.

The application of *MAD* in the equation can be improved to another robust scale estimator so that this robust correlation

coefficient can perform better. Thus, motivated by their work, this study aims to extend the robust correlation coefficient by applying a robust scale estimators namely S_n [9]. This estimator possesses the high breakdown point which is 50% and more efficient under the normality assumption.

III. Research Methodology

This topic focuses on developing a new robust version of correlation coefficient based on the S_n as robust scale estimator. The performance of this proposed procedure will be evaluated based on the value of the correlation, standard errors and average bias from the simulation study.

A. Proposed Robust Correlation Coefficient

The sample correlation coefficient commonly denoted by *r* is the Pearson correlation coefficient as given in Equation (1).

$$r_p = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\left[\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2\right]^{1/2}} \quad (1)$$

Where

x_i = ith observation of variable x
y_i = ith observation of variable y
$\bar{x}$ = the mean of variable x
$\bar{y}$ = the mean of variable y

Another version of robust correlation coefficient has been proposed and known as Median-Product (MP) correlation coefficient [8]. The equation for this robust correlation coefficient is obtained as in Equation (2).

$$r_m = median\left(Q_x \times Q_y\right) \quad (2)$$

Where:

$$Q_x = \frac{\left(x - median\,(x)\right)}{MAD(x)} \quad (3)$$

$$Q_y = \frac{\left(y - median\,(y)\right)}{MAD(y)} \quad (4)$$

Another robust scale estimator can improve the application of *MAD* in the calculation of this coefficient. Therefore, in this study, we propose robust correlation coefficient with the implementation of S_n as the scale estimator. The robust correlation coefficient using this estimator will be named as S_n product correlation coefficient, $(r_{S_n(p)})$. The equation for this coefficient is given in Equations (5) to (8).

$$r_{S_n(p)} = median\,(Q_x \times Q_y) \quad (5)$$

Where:

$$Q_x = \frac{(x - med\,(x))}{S_n(x)} \quad (6)$$

$$Q_y = \frac{(y - med\,(y))}{S_n(y)} \quad (7)$$

$$S_n = c\ med\ \left\{med_i \left| x_i - x_j \right|\right\} \quad (8)$$

Based on the Equation (8), the observation for each *i*, the median for $\left\{\left|x_i - x_j\right|; j = 1, 2, 3, \ldots, n\right\}$ is calculated. This step will provide *n* median that will be used in as the final estimate for S_n. To obtain the S_n value, the *n* median will be multiplied with median and to get the stability and consistency of the S_n, the product of *n* median with median is multiply with the *c* [9]. The value of constant *c* is 1.1926.

B. Data Generation & Sample Size

The performance of the proposed procedure was determined with simulation study by using SAS/IML Version 9.4 [10] generator RANGEN. Random observations of bivariate data will be generating by following the previous study [4], and the ρ is set to 1. Data will be generated based on sample sizes 25, 100 and 400 [8]. The condition of this bivariate data is divided into two. The first condition is the uncontaminated or perfect data and the second condition is contaminated data. For perfect data, the random data is generated with the linear relation of:

$$y_i = 2.0 + 1.0x_i + u_i \quad (9)$$

The observations for x_i is normally distributed along with given; $N(5,1)$. For u_i, the data also normally distributed with $N(0,0.04)$. For the contaminated data where the outlier is present, the data is gradually contaminated with the percentage of 10%, 30%, and 50%. The contaminated data will be generated from the linear relation where y_i is normally distributed with $N(2, 0.04)$, plus x_i is uniformly distributed with parameter [5,10].

C. Simulation Study

To evaluate the performance of the proposed coefficient correlations, 200 datasets will be simulated, and three indicators will be employed that are average of estimates, standard errors and average bias [8]. Three correlation coefficients will be evaluated and compared in this study. Those coefficients are:

1. Pearson correlation coefficient (*r*)
2. Median product correlation coefficient ($r_{m(p)}$)
3. S_n product correlation coefficient ($r_{Sn(p)}$)

The average of estimates will be calculated by finding the average of the value of the correlation coefficients. The closer the value of the correlation coefficients that acquire from the simulation study to 1, the better the performance of the robust correlation coefficients. Following the simulation study in [4], the value of the correlation coefficients will close to 1 when the data is uncontaminated due to the original value of ρ=1. Whereas, the computation procedure for standard error can be calculated by Equation (10).

$$SE = \frac{s}{\sqrt{n}} \quad (10)$$

Where

s = standard deviation
n = sample size

Meanwhile, the process to calculate average bias is by calculating the average of the difference between the outcome value of correlation coefficients in this study with the value of correlation coefficient that had been set in the simulation

which is ρ=1. The smaller the value of average bias, the performance of the robust correlation is better.

IV. Results and Discussion

The performances of the robust correlation coefficient together with the classical correlation coefficient in this study are evaluated through the contamination of the data. The first condition of the simulated data, the data had been set to ρ=1 with the absence of outlier. This simulated data is called perfect data. Meanwhile, the second condition of the simulated data is called contaminated data. The simulated data is gradually contaminated with the percentage of 10%, 30% and 50%.

A. The Performance of the Robust Correlation Coefficient Based on Coefficient Value

The result of the performance of robust correlation coefficient plus classical correlation coefficient in term of the value of coefficients is portraying in Table 1. The performance of the coefficients is better when the value of correlation coefficients is closer to 1. This condition is due to the value of the ρ=1 that have been set for the simulated data. For perfect data, when n=25, 100 and 400, the best result comes out from classical correlation coefficient which is Pearson correlation coefficient (r) with the value of coefficient equal to 0.9990, 0.9990 and 0.9992. Under the same conditions, S_n product correlation coefficient ($rS_{n(p)}$) gives the weakest value of coefficients which are 0.5082, 0.5314 and 0.4842 respectively.

The performance of proposed robust correlation coefficient and the classical correlation coefficient is continued by evaluating the value of the coefficient in the contamination stage of the simulated data. The simulated data is contaminated into three part of percentage which are 10%, 30% and 50%. By referring to Table 1, in the 10% of the contaminated data, all sample sizes show that the outcome of $r_{m(p)}$ correlation coefficient is the best which is 0.8399, 0.6431 and 0.5762, respectively. S_n product correlation coefficient ($rS_{n(p)}$) provides better correlation values compared to the classical correlation.

However, when the simulated data is 30% contaminated, most of the coefficients deteriorate and provides the negatives value of coefficients. Classical correlation performs better compared to the others followed by median product correlation coefficient ($r_{m(p)}$) under all sample sizes. During 50% of contamination of the simulated data, Pearson correlation coefficient (r), still shows good performance compared to robust correlation coefficients. However, for n=400, S_n product correlation coefficient ($rS_{n(p)}$) displays good performance compared to classical correlation, and median product correlation coefficient with r=-0.6293 event thought it produced the less value of the coefficient during the early stage of contamination.

B. Average Bias and Standard Error Value of Classical and Robust Correlation Coefficient for Perfect Data

The performances of the robust correlation coefficient and the classical correlation coefficient also measured regarding average bias and standard error. The performance of the coefficient is better when the value of average bias is low. A good coefficient is distinguished when the value of the standard error is closer to 0. The average bias and standard error for the three coefficients are displayed in Table 2 to Table 5. Table 2 displays the average bias and standard error for perfect data.

Table 1
The Coefficient Value Under Perfect and Contaminated Data (n=25,100,400 and ρ=1)

Correlation Coefficients				
Data	Sample sizes	r	$r_{m(p)}$	$rS_{n(p)}$
Perfect Data	25	0.9990	0.9811	0.5082
	100	0.9990	0.9220	0.5314
	400	0.9992	0.9800	0.4842
Contaminated Data (10%)	25	-0.1386	0.8399	0.4097
	100	-0.2763	0.6431	0.3175
	400	-0.1712	0.5762	0.2656
Contaminated Data (30%)	25	-0.6885	-0.2490	-0.1025
	100	-0.5289	-0.0540	-0.0076
	400	-0.4846	-0.0540	-0.0243
Contaminated Data (50%)	25	-0.7320	-0.4406	-0.2235
	100	-0.6102	-0.4718	-0.4266
	400	-0.5892	-0.0977	-0.6293

Table 2
The Average Bias and Standard Error of Classical Correlation and Robust Correlation in the Perfect Data (n=25,100,400)

Correlation coefficient	n=25		n=100		n=400	
	Ave. Bias	Std error	Ave. Bias	Std error	Ave. Bias	Std error
r	0.0008	0.0000	0.0008	0.0000	0.0008	0.0000
$r_{m(p)}$	0.0055	0.0024	0.0032	0.0016	0.0024	0.0011
$r_{Sn(p)}$	0.0025	0.0025	0.0023	0.0027	0.0026	0.0024

By referring to Table 2, the classical correlation coefficient which is Pearson correlation coefficient yields the best value of average bias and standard error for perfect data. When n = 25, 100 and 400, the value of average bias for Pearson correlation coefficient is the lowest which is 0.0008. Meanwhile, the value of standard error for Pearson correlation coefficient (r) also the lowest which is 0.000 for all sample sizes. S_n product correlation coefficient ($rS_{n(p)}$) give the lower value of average bias compared to median product correlation coefficient ($r_{m(p)}$) under small and medium sample sizes and on par with each other under large sample size.

The performance of the classical correlation coefficient with robust correlation coefficient is continued being observed under the contamination of the data. Table 3 presents the result of average bias and the standard error of the classical correlation coefficient and robust correlation coefficient under 10% contamination. Based on Table 3, median product correlation coefficient ($r_{m(p)}$) gives the lowest value of average bias for all sample sizes. The value of average bias is 0.5092, 0.4286 and 0.4217. Meanwhile, Pearson correlation coefficient produced the largest average bias for all sample sizes. In the meantime, S_n product correlation coefficient ($rS_{n(p)}$) gives the lowest value of standard error compared to the other two coefficients.

The performance of average bias and standard error for classical correlation coefficient and robust correlation coefficient are compared with 30% contamination of the data as shown in Table 4. S_n product correlation coefficient ($rS_{n(p)}$) provides the lowest value of average bias and standard error for all conditions. Despite that, the performance of Pearson correlation coefficient in terms of average bias and standard error is still the worst.

Table 3
The Value of Average Bias and Standard Error of Classical Correlation and Robust Correlation in the 10% Contamination of Data (n=25,100,400)

Correlation coefficient	*n*=25		*n*=100		*n*=400	
	Ave. Bias	Std error	Ave. Bias	Std error	Ave. Bias	Std error
r	1.2037	0.0148	1.1201	0.0078	1.0991	0.0040
$r_{m(p)}$	0.5092	0.0118	0.4286	0.0060	0.4217	0.0026
$r_{Sn(p)}$	0.8148	0.0048	0.7622	0.0030	0.7552	0.0014

Table 4
The Value of Average Bias and Standard Error of Classical Correlation and Robust Correlation in the 30% Contamination of Data (n=25,100,400)

Correlation coefficient	*n*=25		*n*=100		*n*=400	
	Ave. Bias	Std error	Ave. Bias	Std error	Ave. Bias	Std error
r	1.5348	0.0082	1.4876	0.0051	1.4880	0.0024
$r_{m(p)}$	1.0995	0.0060	1.0580	0.0036	1.0679	0.0019
$r_{Sn(p)}$	1.0387	0.0023	1.0236	0.0015	1.0278	0.0008

During the 50% contamination of data, the performance of classical correlation coefficient together with robust correlation coefficient is poor in term of average bias and standard error. 50% contamination of data means that half of the data is an outlier and affecting the other half of clean data. Thus, the result of average bias and standard error for the three coefficients in this study is unsteady. Table 5 indicates the result of three coefficients based on average bias and standard error indicators.

Table 5
The Value of Average Bias and Standard Error of Classical Correlation and Robust Correlation in the 50% Contamination of Data (n=25,100,400)

Correlation coefficient	*n*=25		*n*=100		*n*=400	
	Ave. Bias	Std error	Ave. Bias	Std error	Ave. Bias	Std error
r	1.5943	0.0077	1.5838	0.0037	1.5749	0.0018
$r_{m(p)}$	1.4245	0.0151	1.5468	0.0154	1.6771	0.0108
$r_{Sn(p)}$	1.2173	0.0073	1.4531	0.0039	1.5064	0.0033

Referring to the Table 5, S_n product correlation coefficient ($rS_{n(p)}$) has the lowest value of average bias while Pearson correlation coefficient has the largest average bias. Regarding the standard error, $rS_{n(p)}$ gives the lowest value of standard error for n=25 while r has the lowest value of standard error for n=100 and 400.

V. CONCLUSION

Real datasets usually contain a fraction of outliers and other contaminations. The classical correlation coefficient such as Pearson's product moment correlation coefficient r is much affected by the outliers and often gives misleading results. Robust methods are designed to consider the majority of the data rather than all the data. Therefore, robust methods give reasonable results even when data contain a fraction of outliers. To achieve robustness and computational efficiency, we proposed a new robust estimator of correlation. The classical estimator of correlation uses non-robust estimator mean and standard deviation as the building blocks. In this study, we construct the new robust correlation coefficient by replacing these non-robust estimators with their robust counterpart S_n.

Under the condition of perfect data, classical correlation performs the best. However, its performance becomes worst when data are contaminated. Regarding the correlation value, the performance of S_n product correlation coefficient is less compared to median product correlation coefficient. However, regarding average bias and standard error, S_n product correlation coefficient performs better compared to the others in most of the condition under study.

ACKNOWLEDGEMENT

We earnestly acknowledge the Universiti Utara Malaysia for the financial support under Universiti Grant Scheme (Code S/O 13377) and RIMC for facilitating the management of the research.

REFERENCES

[1] F. E. Grubbs, "Procedures for detecting outlying observations in samples", *Technometrics*, vol, 11, no. 1, pp. 1-21, 1969.
[2] P. J. Huber, "Robust estimation of a location parameter", *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73-101, 1964.
[3] J. W. Tukey, "A survey of sampling from contaminated distributions", *Contributions to Probability and Statistics*, vol. 2, pp. 448-485, 1960.
[4] M. B. Abdullah, "On a robust correlation coefficient", *The Statistician*, pp. 455-460, 1990.
[5] E. B. Niven and C. V. Deutsch, "Calculating a robust correlation coefficient and quantifying its uncertainty", *Computers & Geosciences*, vol. 40, pp. 1-9, 2012.
[6] J. Kim and J. A. Fessler, "Intensity-based image registration using robust correlation coefficients", *Medical Imaging, IEEE Transactions on*, vol. 23, no. 11, pp. 1430-1444, 2004.
[7] G. L. Shevlyakov, "On robust estimation of a correlation coefficient", *Journal of Mathematical Sciences*, vol. 83, no. 3, pp. 434-438, 1997.
[8] A. Z. M. Shafiullah and J. A. Khan, "A new robust correlation estimator for bivariate data". *Bangladesh Journal of Scientific Research*, vol. 24, no. 2, pp. 97-106, 2012.
[9] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation". *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273-1283, 1993.